



**You have downloaded a document from
RE-BUS
repository of the University of Silesia in Katowice**

Title: Objectivisation via subjectivisation in annotated corpus linguistics

Author: Ireneusz Kida

Citation style: Kida Ireneusz. (2011). Objectivisation via subjectivisation in annotated corpus linguistics. "Studia Methodologica" (T. 33 (2011), s. 34-37).



Uznanie autorstwa - Licencja ta pozwala na kopiowanie, zmienianie, rozprowadzanie, przedstawianie i wykonywanie utworu jedynie pod warunkiem oznaczenia autorstwa.



UNIwersYTET ŚLĄSKI
W KATOWICACH



Biblioteka
Uniwersytetu Śląskiego



Ministerstwo Nauki
i Szkolnictwa Wyższego

Ireneusz KIDA

OBJECTIVISATION VIA SUBJECTIVISATION IN ANNOTATED CORPUS LINGUISTICS

In my article I would like to suggest how to annotate the linguistic material in order to reflect the dynamic and changeable nature of language in the analysis of word order configurations. The words “dynamic and changeable” mean that languages constantly change in time, and therefore they are never dead. This observation does not only refer to the spoken language but also to the written language.

AMBIVALENT NATURE OF SOME ANCIENT ENGLISH SENTENCES

It is logical to think that the development from parataxis to hypotaxis, or from hypotaxis to parataxis, is not an abrupt one and that there is always a transition stage where both paratactic and hypotactic sentences coexist. According to Jucker (1991:203) “there must be one or possibly several intermediate stages between true parataxis and true hypotaxis and that there are constructions that are neither clearly paratactic nor clearly hypotactic but somewhere in-between. In most cases this development will have been not so much a matter of discrete steps, but rather a gradual movement, which makes it difficult to ascertain the exact status of a construction at any one time”. Although Ancient English achieved an advanced stage of hypotaxis, we can often have problems with the classification of some sentences. As Baugh and Cable (1993:66) indicate, “there are clear differences in our modern perceptions of Ancient English written in paratactic style and Ancient English written with many subordinate sentences. The problem is in determining whether a particular sentence is independent or subordinate, because the words that do the subordinating are often ambiguous. The Ancient English *þa* at the beginning of a sentence can be either an adverb translated ‘when’ and indicating an independent sentence, or a subordinating conjunction translated ‘when’ and introducing a dependent sentence. Similarly, *þær* can be translated as ‘there’ or ‘where’, *þonne* as ‘then’ or ‘when’, *swa* as ‘so’ or ‘as’, *ær* as ‘formerly’ or ‘ere’, *siððan* as ‘afterward’ or ‘since’, *nu* as ‘now’ or ‘now that’, *þeah* as ‘nevertheless’ or ‘though’ and *forðam* as ‘therefore’ or ‘because’” (p.66). They also say that in each pair the first word is an adverb, and the style that results from choosing it is a simpler style with shorter sentences, whereas the choice of the second word results in longer sentences with more embedded sentences. Moreover, they note that current research in Ancient English syntax aims to understand the use of these ambiguous subordinate pronouns and adverbs. The conclusions that emerge will affect our modern perception of the sophistication of Ancient English writing in verse and prose. They also note that we should be especially cautious about imposing modern notions that equate hypotaxis with sophistication and parataxis with primitiveness until we know more about the full range of syntactic possibilities in Ancient English. Ongoing research in this subject promises to revise our ideas of the grammatical, semantic, and rhythmic relationships in Ancient English verse and prose”²³ (Baugh and Cable 1993:67). Mitchell (1985:§1879) warns us that it may be anachronistic to impose modern categories resulting from our translations into words like ‘then’ and ‘when’, “implying that the choice was simply between a subordinate sentence and

23 For more information on this issue see Mitchell (1985) Mitchell (1988), Mitchell and Robinson (2007), Blake (1992), Denison (1993), Fischer, Kemenade, Koopman, and Wurff (2000), Hogg (1992), Kohonen (1978), Molencki (1997), Pintzuk (1993, 1995).

an independent sentence in the modern sense of the words". Moreover, he says that one should be especially careful about employing modern notions that, on the one hand, equate parataxis with primitiveness and, on the other, hypotaxis with sophistication until more research is done about the full range of the syntactic possibilities that Ancient English had at its disposal. Baker (2003:29) observes that some linguists claim that Ancient English literature is generally characterised by parataxis, but it is not so, because it is only some Ancient English works, such as the *Anglo-Saxon Chronicle* for example, that tend to be paratactic, whereas other works, like King Alfred's Preface to his translation of Gregory's *Pastoral Care* for example, are characterised by hypotaxis. He further says that in Ancient English it can be difficult to tell independent sentences from subordinate sentences, and because of that it is a matter of some controversy how paratactic or hypotactic Ancient English was in fact.

The prevalent existence of ambiguity in language poses a serious problem for the constructors of annotated corpora. Baker et al. (2006:10) note that "in corpus annotation, in cases where there is a choice of two potential annotations at one point in the text, it is not always possible to make a clear-cut decision. [...] In some cases a portmanteau annotation can be given in order to address the ambiguity. In other words, examining more of the surrounding context may help to solve the problem. However, in extremely ambiguous cases, the corpus builder may have to make a decision one way or the other. If this approach is taken, then the decision would need at least to be applied with consistency throughout the corpus. Pala et al. (1997) say that the most reasonable way of building large annotated corpora is via an automatic annotation of the texts by means of computer programs. However, they add that "natural languages have rather complex sentence and therefore it is no surprise that the attempts to process them by the simple deterministic algorithms do not always give satisfactory results. The result is that the present annotating programs are not able to give fully reliable results and there are many ambiguities in their output" (p.523).

AN ANNOTATED CORPUS WITH PH SENTENCES

In an annotated corpus that does not take into account PH sentences, all the sentences are treated as unambivalent and they are therefore given unambivalent tags. Generally speaking, in such a corpus there are a fixed number of unambivalent main sentences and a fixed number of unambivalent dependent sentences, and the number of both kinds of sentences does not change once the corresponding tags have been added. For example, in such a corpus the Modern English main sentence *John reads books* will be tagged as a main sentence:

John reads books
 +=+S+V+DO+,²⁴

On the other hand, the sentence *John reads books* in the utterance *I know that John reads books* will be treated as a dependent sentence and will be assigned a corresponding code:

I know that John reads books
 +=+s+V+,+con*+S+V+DO+²⁵,

However, in my annotated corpus that takes PH sentences into account, apart from the fact that there are a fixed number of both unambivalent main and dependent sentences, there are also a certain number of PH sentences whose status is ambivalent and therefore they are annotated in two ways. Let us take the following example from the ASC E for illustration:

[000500 (0.10)E] We witan oþer egland her be easton **þer** ge magon eardian gif ge willað ...

H1+(**con***)+s+V+inf+,
 +=+s+V+DO+X+,**PH1**}{=x/[**con***,+con*+s+V+...
 P1+(=)+x+s+V+inf+,

H1 'We know another island here to the east **where** you may dwell, if you will'²⁶

24 The sign = stands for a main sentence, *S* stands for a nominal subject, *V* stands for a finite verb, and *DO* stands for a nominal direct object. Moreover, every sign is separated by a +.

25 The sign * stands for a dependent sentence, *con* stands for a dependent sentence connective, and *s* stands for a pronominal subject. Moreover, every sign is separated by a +.

26 The translation of the Ancient English example comes from <http://www.britannia.com/history/docs/asintro2.html>.

P1 ‘We know another island here to the east; **there** you may dwell, if you will’

It can be noticed that in this utterance there is one unambivalent main sentence (*We witan oþer egland her be easton*), one unambivalent dependent sentence (*gif ge willað*), as well as one ambivalent PH sentence, namely *þer ge magon eardian*. Since the PH element *þer* is ambivalent (because, on the one hand, it can be translated as ‘there’ and, on the other, as ‘where’), the whole sentence immediately following it is ambivalent. This ambivalent PH sentence therefore has two codes, one for a main sentence, namely $P1+(=)+x+s+V+inf+$, and the other for a dependent sentence, namely $H1+(con^*)+s+V+inf+$. In the main sentence, the adverbial *þer* is treated as an ordinary pronominal adverbial and is assigned the tag *x*. In the dependent sentence, on the other hand, it is treated as a dependent sentence connective and is assigned the tag *con**, and the whole sentence turns into a V2 sentence.

HOW TO USE A PH CORPUS?

Having at his disposal an annotated corpus, in which ambivalent PH sentences are tagged in two ways, when the end-user comes across the ambivalent PH sentence *þer ge magon eardian*, he can analyse it from two perspectives, namely from the paratactic perspective and from the hypotactic perspective, and then compare the results. However, in case he is convinced of a hypotactic status of this sentence, and would like to analyse it only from the hypotactic point of view, all he needs to do is to remove the round brackets from $H1+(con^*)+s+V+inf+$ in order to obtain $H1+con^*+s+V+inf+$ and make it possible for the computer to follow the hypotactic track $...con^*+...$; all the unambivalent sentences, both main and dependent, are tagged without the use of round brackets for the computer to read only the sequences of tags without round brackets; the round brackets are used when the sentences are as if in a state of waiting for being analysed by the computer. By this procedure, namely by removing the brackets, the end-user would obtain the following sequence:

[000500 (0.10)E] We witan oþer egland her be easton **þer ge magon eardian** gif ge willað, ...
 $H1+con^*+s+V+inf+$,
 $+s+V+DO+X+,PH1\} [=x/[con^*,+con^*+s+V+...$
 $P1+(=)+x+s+V+inf+$,

On the other hand, given that the end-user is convinced of a paratactic status of the sentence in question and would like to analyse it from the paratactic point of view, he needs to remove the round brackets from $P1+(=)+x+s+V+inf+$ in order to obtain $P1+=+x+s+V+inf+$ and make it possible for the computer to follow the paratactic track $...+=+...$. By performing this operation, namely removing the brackets, he will obtain the following configuration:

[000500 (0.10)E] We witan oþer egland her be easton **þer ge magon eardian** gif ge willað, ...
 $H1+(con^*)+s+V+inf+$,
 $+s+V+DO+X+,PH1\} [=x/[con^*,+con^*+s+V+...$
 $P1+=+x+s+V+inf+$,

In this way the computer will follow the paratactic track, because the hypotactic track stays intact because of the round brackets that block the hypotactic analysis.

Therefore, depending on the decision of the end-user, the computer will follow either the hypotactic track or the paratactic one, and the PH sentences will be analysed either as dependent or main respectively, together with the unambivalent main and dependent sentences of the corpus. The end-users analysing the ambivalent PH sentences do not have to analyse all of them either as main or dependent. They can consider only some of them as main and the remaining ones as dependent. Therefore, depending on the end-users the data will be slightly different, and the more end-users there are, the more subjective choices there will be made; it is what we term *objectivisation via subjectivisation*. In other words, the data obtained in the analysis of the ASC annotated corpus will be objectivised on the basis of the subjective choices of various end-users with respect to the establishment of the status of PH sentences. And the more choices there are, the better, because the probability that given PH sentences are main/dependent will be closer to the reality. Moreover, the subjective choices can only be made within the boundaries of the ambivalence corridor²⁷ because the non-PH sentences, whose status is established *a priori*, are

27 A corridor of ambivalence is the maximum number of main and dependent sentences, with their word order configurations, that can be obtained from given PH sentences.

unambivalently main or dependent.

CONCLUSION

However, the idea of *objectivisation via subjectivisation* can be treated as optional. Nevertheless, I think that the end-users should have the possibility of deciding about the status of PH sentences because they might not be satisfied with the maximum ranges of the ambivalence corridors that we propose, in which case they would be able to create their own corridors of ambivalence within the corridors of ambivalence that we establish. Nevertheless, in our analysis we are concerned only with the maximum ranges of the ambivalence corridors, and we do not take into account any intermediate stages, it means that when we treat all the PH sentences either as main or as dependent. The dual analysis of PH sentences has an influence upon the general state of word order configurations in Ancient English texts. In order to observe the influence, on the basis of different kinds of ambivalent PH sentences, which are annotated from two different perspectives, we establish the maximum ranges of ambivalence corridors in both manuscripts and then incorporate them into the data obtained from the analysis of unambivalent main and dependent sentences. In this way it can be seen that the general state of word order configurations in Ancient English changes depending on whether the PH sentences are approached from the paratactic point of view, in which case they are analysed as main sentences, or from the hypotactic point of view, in which case they are analysed as dependent sentences. The problem of PH sentences should not be neglected in the construction of annotated corpora of Ancient English because their widespread existence cannot be denied. The two different approaches, paratactic and hypotactic, to these sentences can lead to serious fluctuations in word order configurations because depending on whether the PH sentences are approached from the paratactic point of view or from the hypotactic point of view, the general picture of the individual word order configurations in the main and in the dependent sentences of a given text will vary.